LightStone

第5回　マルチレベル分析とSEM

SEM の最終回です. 今回は回帰モデルに入れ子型の階層構造を仮定するマルチレベル分析と SEM の考え方を合わせて利用するコマンドを紹介します. すなわち, 観測できない潜在変数による階層構造を考えてモデルを作成することになります. Stata 15 の [SEM] マニュアルにある example 30g を利用して解説を行います. ただし, ここではマルチレベル分析の手法に関する説明は行いませんのでご了承ください.

2 レベルの計測モデル

データをダウンロードして内容を確認します.

```
. use http://www.stata-press.com/data/r15/gsem_cfa,clear
```

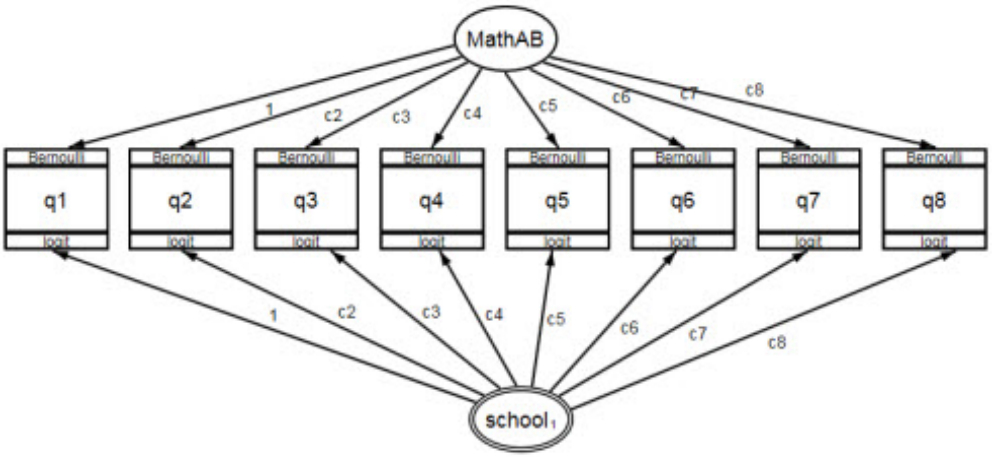
```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
school	500	10.5	5.772056	1	20
id	500	50681.71	29081.41	71	100000
q1	500	.506	.5004647	0	1
q2	500	.394	.4891242	0	1
q3	500	.534	.4993423	0	1
q4	500	.424	.4946852	0	1
q5	500	.49	.5004006	0	1
q6	500	.434	.4961212	0	1
q7	500	.52	.5001002	0	1
q8	500	.494	.5004647	0	1
att1	500	2.946	1.607561	1	5
att2	500	2.948	1.561465	1	5
att3	500	2.84	1.640666	1	5
att4	500	2.91	1.566783	1	5
att5	500	3.086	1.581013	1	5
test1	500	75.548	5.948653	55	93
test2	500	80.556	4.976786	65	94
test3	500	75.572	6.677874	50	94
test4	500	74.078	8.845587	43	96

これは人工的に作成した擬似データで, 500 人の生徒の数学能力を示すものです. 変数 q1 から q8 は 0 と 1 を使って単純に数学の設問に対する正解と不正解の情報を示すものです.

2 レベルモデルのフィット

生徒はそれぞれの学校に属しており, 変数 `school` は学校の ID です. 潜在変数としては生徒単位の数学能力以外に, 学校単位での教育効果というものが考えられます. イメージとして次のようなものになります.



下側にある二重線による school_1 は今回, 新たに登場したオブジェクトです. このオブジェクトの意味は学校単位で異なる潜在変数を意味しています.

仮に, school ごとの潜在変数を考えない場合は次のコマンドを実行します.

```
. gsem (MathAb -> q1-q8), logit
```

(結果は省略します)

これに対し, school レベルの効果を考える場合は次に示すように外生変数として `M1[school]` を追加します. ここでは決まりとして大文字の M を利用します.

```
. gsem (MathAb M1[school] -> q1-q8), logit
```

```
Fitting fixed-effects model:
Iteration 0:   log likelihood = -2750.3114
Iteration 1:   log likelihood = -2749.3709
Iteration 2:   log likelihood = -2749.3708
Refining starting values:
Grid node 0:   log likelihood = -2649.0033
Fitting full model:
Iteration 0:   log likelihood = -2649.0033  (not concave)
Iteration 1:   log likelihood = -2645.0613  (not concave)
Iteration 2:   log likelihood = -2641.9755  (not concave)
Iteration 3:   log likelihood = -2634.3857
Iteration 4:   log likelihood = -2631.1111
Iteration 5:   log likelihood = -2630.7898
Iteration 6:   log likelihood = -2630.2477
Iteration 7:   log likelihood = -2630.2402
Iteration 8:   log likelihood = -2630.2074
Iteration 9:   log likelihood = -2630.2063
Iteration 10:  log likelihood = -2630.2063
Generalized structural equation model           Number of obs     =        500
Response       : q1
Family         : Bernoulli
Link           : logit
Response       : q2
Family         : Bernoulli
Link           : logit
```

```
Response       : q3
Family         : Bernoulli
Link           : logit
Response       : q4
Family         : Bernoulli
Link           : logit
Response       : q5
Family         : Bernoulli
Link           : logit
Response       : q6
Family         : Bernoulli
Link           : logit
Response       : q7
Family         : Bernoulli
Link           : logit
Response       : q8
Family         : Bernoulli
Link           : logit
Log likelihood = -2630.2063
 ( 1)  [q1]M1[school] = 1
 ( 2)  [q2]MathAb = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
q1						
M1[school]	1	(constrained)				
MathAb	2.807515	.9468682	2.97	0.003	.9516878	4.663343
_cons	.0388021	.1608489	0.24	0.809	-.276456	.3540602
q2						
M1[school]	.6673925	.3058328	2.18	0.029	.0679712	1.266814
MathAb	1	(constrained)				
_cons	-.4631159	.1201227	-3.86	0.000	-.698552	-.2276798
q3						
M1[school]	.3555867	.3043548	1.17	0.243	-.2409377	.9521111
MathAb	1.455529	.5187786	2.81	0.005	.4387416	2.472316
_cons	.1537831	.1070288	1.44	0.151	-.0559894	.3635556
q4						
M1[school]	.7073241	.3419273	2.07	0.039	.037159	1.377489
MathAb	.8420897	.3528195	2.39	0.017	.1505762	1.533603
_cons	-.3252735	.1202088	-2.71	0.007	-.5608784	-.0896686
q5						
M1[school]	.7295553	.3330652	2.19	0.028	.0767595	1.382351
MathAb	2.399529	.8110973	2.96	0.003	.8098079	3.989251
_cons	-.0488674	.1378015	-0.35	0.723	-.3189533	.2212185
q6						
M1[school]	.484903	.2844447	1.70	0.088	-.0725983	1.042404
MathAb	1.840627	.5934017	3.10	0.002	.6775813	3.003673
_cons	-.3139302	.1186624	-2.65	0.008	-.5465042	-.0813563
q7						
M1[school]	.3677241	.2735779	1.34	0.179	-.1684787	.903927
MathAb	2.444023	.8016872	3.05	0.002	.8727449	4.015301
_cons	.1062164	.1220796	0.87	0.384	-.1330552	.3454881
q8						

M1[school]	.5851299	.3449508	1.70	0.090	-.0909612	1.261221
MathAb	1.606287	.5367614	2.99	0.003	.5542541	2.65832
_cons	-.0261962	.1189835	-0.22	0.826	-.2593995	.2070071
var(M1[school])	.2121216	.1510032			.052558	.8561121
var(MathAb)	.2461246	.1372513			.0825055	.7342217

推定結果

1. M1[school] の分散推定値は 0.21
2. M1[school] の値の意味をどう解釈すれば良いでしょうか?MathAb の分散推定値は 0.25 なので, 大きさはほぼ同じと考えられます. しかし, 実際の調査により入手できる現実のデータでは個人の数学能力 MathAb の方が, 学校の分散推定値よりも大きいケースが殆どです. したがって, 次のようにモデルを変更します.

分散コンポーネントモデルのフィット

先の推定例では潜在変数 MathAB と学校単位の教育の特徴は独立としてモデル化しました. しかし, 次に示すように学校ごとの教育の特徴が数学能力にも影響を与えていると考えることもできます.

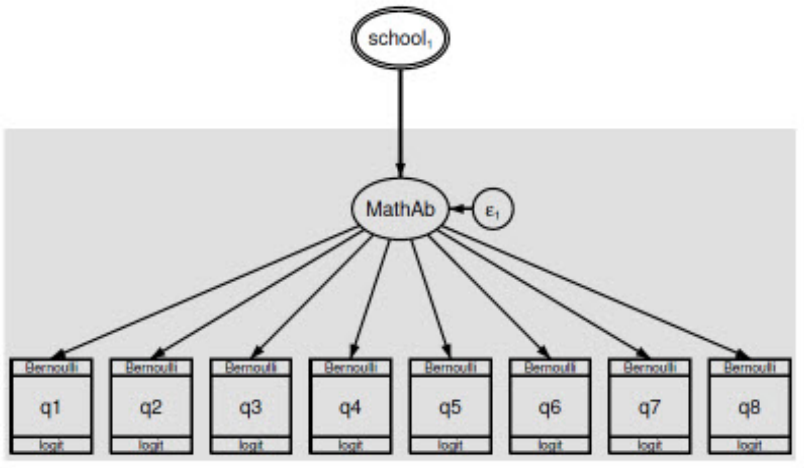
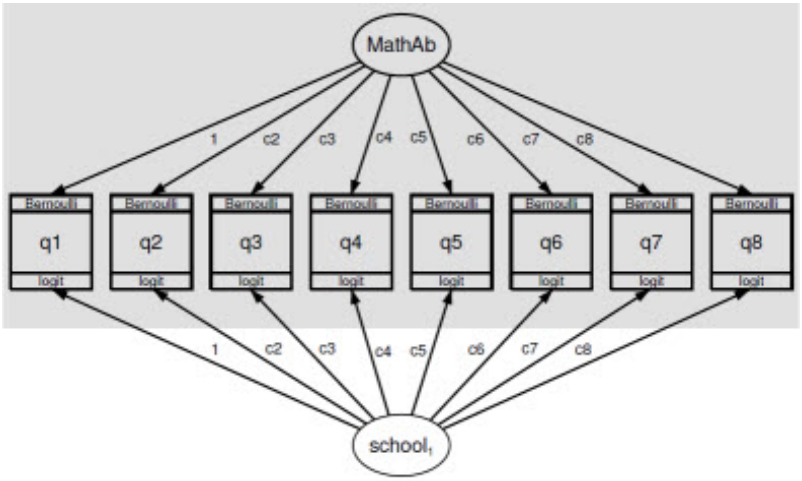


図 1. マルチレベルと SEM を利用したイメージ

実際に Stata の sem builder でこのようなパス図を描くことは仕様上, 出来ません. このような場合は, 代わりに次のようなパス図を作成し, 制約を課すというステップを用います.



2 つのモデルは一見, 異なるように見えますが, モデルに制約を掛けることで同一のモデルになります.

```
. gsem (MathAb M1[school] ->
>       q1@1 q2@c2 q3@c3 q4@c4 q5@c5 q6@c6 q7@c7 q8@c8), logit


Fitting fixed-effects model:
Iteration 0:   log likelihood = -2750.3114
Iteration 1:   log likelihood = -2749.3709
Iteration 2:   log likelihood = -2749.3708
Refining starting values:
Grid node 0:   log likelihood = -2642.8248
Fitting full model:
Iteration 0:   log likelihood = -2651.7239  (not concave)
Iteration 1:   log likelihood = -2644.4937
Iteration 2:   log likelihood =   -2634.92
Iteration 3:   log likelihood = -2633.9336
Iteration 4:   log likelihood = -2633.5924
Iteration 5:   log likelihood = -2633.5922
Generalized structural equation model           Number of obs     =        500
Response       : q1
Family         : Bernoulli
Link           : logit
Response       : q2
Family         : Bernoulli
Link           : logit
Response       : q3
Family         : Bernoulli
Link           : logit
Response       : q4
Family         : Bernoulli
Link           : logit
Response       : q5
Family         : Bernoulli
Link           : logit
Response       : q6
Family         : Bernoulli
Link           : logit
Response       : q7
```

```
Family         : Bernoulli
Link           : logit
Response       : q8
Family         : Bernoulli
Link           : logit
Log likelihood = -2633.5922
 ( 1)  [q1]M1[school] = 1
 ( 2)  [q1]MathAb = 1
 ( 3)  [q2]M1[school] - [q2]MathAb = 0
 ( 4)  [q3]M1[school] - [q3]MathAb = 0
 ( 5)  [q4]M1[school] - [q4]MathAb = 0
 ( 6)  [q5]M1[school] - [q5]MathAb = 0
 ( 7)  [q6]M1[school] - [q6]MathAb = 0
 ( 8)  [q7]M1[school] - [q7]MathAb = 0
 ( 9)  [q8]M1[school] - [q8]MathAb = 0
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
q1						
M1[school]	1	(constrained)				
MathAb	1	(constrained)				
_cons	.0385522	.1556214	0.25	0.804	-.2664601	.3435646
q2						
M1[school]	.3876281	.1156823	3.35	0.001	.1608951	.6143612
MathAb	.3876281	.1156823	3.35	0.001	.1608951	.6143612
_cons	-.4633143	.1055062	-4.39	0.000	-.6701028	-.2565259
q3						
M1[school]	.4871164	.1295515	3.76	0.000	.2332001	.7410328
MathAb	.4871164	.1295515	3.76	0.000	.2332001	.7410328
_cons	.1533212	.1098068	1.40	0.163	-.0618962	.3685386
q4						
M1[school]	.3407151	.1058542	3.22	0.001	.1332446	.5481856
MathAb	.3407151	.1058542	3.22	0.001	.1332446	.5481856
_cons	-.3246936	.1011841	-3.21	0.001	-.5230108	-.1263763
q5						
M1[school]	.8327426	.1950955	4.27	0.000	.4503624	1.215123
MathAb	.8327426	.1950955	4.27	0.000	.4503624	1.215123
_cons	-.0490579	.1391324	-0.35	0.724	-.3217524	.2236365
q6						
M1[school]	.6267415	.1572247	3.99	0.000	.3185868	.9348962
MathAb	.6267415	.1572247	3.99	0.000	.3185868	.9348962
_cons	-.3135398	.1220389	-2.57	0.010	-.5527317	-.074348
q7						
M1[school]	.7660343	.187918	4.08	0.000	.3977219	1.134347
MathAb	.7660343	.187918	4.08	0.000	.3977219	1.134347
_cons	.1039102	.1330652	0.78	0.435	-.1568927	.3647131
q8						
M1[school]	.5600833	.1416542	3.95	0.000	.2824462	.8377203
MathAb	.5600833	.1416542	3.95	0.000	.2824462	.8377203
_cons	-.0264193	.1150408	-0.23	0.818	-.2518951	.1990565
var(M1[school])	.1719347	.1150138			.0463406	.6379187
var(MathAb)	2.062489	.6900045			1.070589	3.973385

- 各推定式で MathAb の係数と M1[school] のパスの係数が等しいことが分かります.
- M1[school] と MathAb の分散推定値は 0.17 と 2.06 で大きく異なります.
- 学校ごとの効果に比べ, 生徒毎の数学能力の方がテストの結果に大きく影響することが分かります.

SEM ビルダーで 2 レベルモデルを作成する

SEM ビルダーを使って, 制約のあるモデルを作図します. 手順として最初に無制約のモデルを推定します.

1. **統計**>**(SEM) 構造方程式モデリング**> **モデル構築/推定**と操作します.
2. 

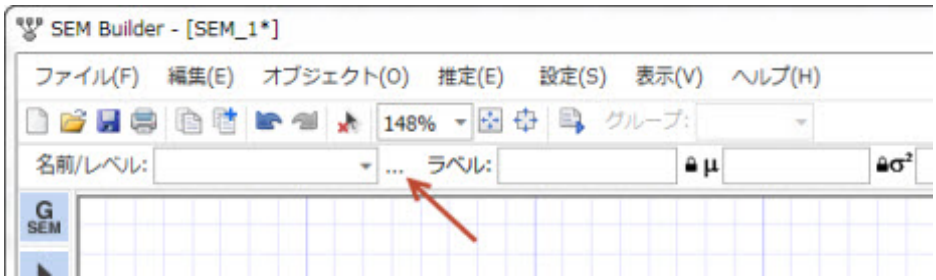
ボタンをクリックし, SEM ビルダーを gsem モードに変更します.
3. 計測可能な MathAb の構成要素を作図します.

   計測要素を追加する  ボタンをクリックします. そして画面の上から 1/4 あたりの箇所をクリックします. そして表示される測定成分のダイアログで次のように操作します.

   (a) 潜在変数の項目でグループ変数名を MathAb と入力します.
   (b) 測定変数の項目に変数として q1-q8 と入力します.
   (c) 同じダイアログで「測定変数を一般化する」の項目をチェックします.
   (d) 選択肢として Bernoulli,Logit を選択します.
   (e) 測定の向きは下方向にします.
   (f) OK ボタンをクリックします.

   必要に応じてパス図の位置調整を行います.
4. school レベルの潜在変数を作図します.

   (a) マルチレベルの潜在変数を追加するアイコン  をクリックします. そして潜在変数 MathAb と対応する下側の位置をクリックします.
   (b) ツールバーにある次のボタンをクリックします.



(c) ダイアログで入れ子のレベル数と変数を設定します.

(d) ベース名は M1 とします.

(e) OK ボタンをクリックします.

5. マルチレベルの潜在変数 `school` に対する因子負荷量のパスを作図します.

(a) パスを追加する アイコンをクリックします.

(b) マルチレベルの潜在変数の school_1 から計測可能な変数 q1-q8 へそれぞれパスを追加します.

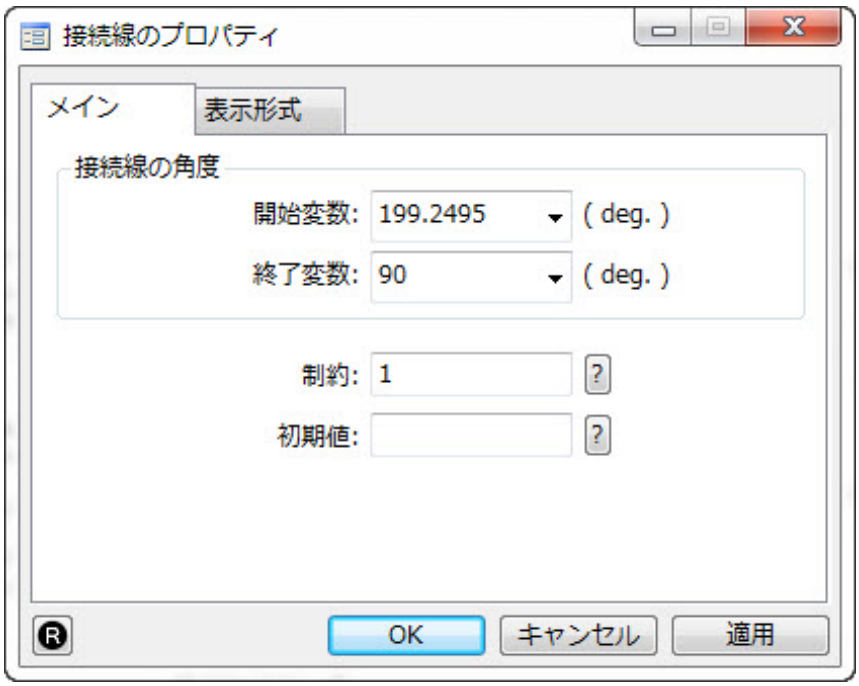
6. 推定ボタン  をクリックし, GSEM によるモデル推定を実行します.

7. このモデルに制約を掛けて目的とする 2 レベルの SEM モデルを推定します.

(a) **推定**> **推定結果をクリア**と操作します.

(b) オブジェクトを選択する  ボタンをクリックします.

(c) `MathAb` から `q1` に向かうパスをダブルクリックして次に示すダイアログを表示し, 制約のテキストボックスに 1 と入力します.



(d) 同じ操作を $\texttt{school}_1$ から q1 に向かうパスについても行います.

(e) q2 のパスをダブルクリックしてそれぞれ制約を課します. ただし, テキストボックスには c2 と入力します.

(f) この操作を q3 から q8 まで繰り返します. もちろん, テキストボックスには c3 から c8 を入力します.

8. 再び, 推定ボタンをクリックし, GSEM によるモデル推定を実行します.

以上が SEM においてマルチレベル分析を利用した推定例になります. ここで考え方を整理するために, 数式を用いて SEM とマルチレベルのモデルを表現してみます.

- 話を簡単にするために, 数学の問題は 1 問しかない状況を想定し, SEM を考えると

$$q_{ij} = \alpha_0 + \beta_0 \,\mathrm{MathAb} + e_{ij}$$

MathAb は計測できない潜在変数です. j は学校の ID, i はその学校に属する生徒の ID.q は数学の試験の得点です.

- 次に学校ごとに, 生徒の数学の学力に及ぼす計測できない要因があると考えます.

$$q_{ij} = \alpha_0 + \beta_0 \,\mathrm{MathAb} + u_{0j} + e_{ij}$$

u_{0j} は学校単位の変量効果です.

- マルチレベル分析では固定効果として計測可能な変数を利用するという決まりがありましたが, SEM を利用することで潜在変数を固定効果部分に用いることができるようになります.

マルチプルグループに対応したワイブルサバイバルモデル

gsem に新たに用意された group オプションの用法を [SEM] マニュアルの example 49g を用いて解説します. ここでの目的は gsem コマンドと新しいコマンドオプション group を利用してワイブル分布に従うパラメトリックサバイバルモデルを推定することです.

最初にサンプルデータを読み込みます.

```
. webuse gsem_cancer,clear
```

データの内容を確認します.

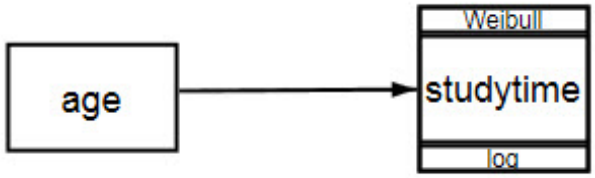
```
. describe
```

```
Contains data from http://www.stata-press.com/data/r15/gsem_cancer.dta
  obs:            48                          Patient Survival in Drug Trial
 vars:             4                          16 Jan 2017 15:41
 size:           192                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
studytime       byte    %8.0g                 Months to death or end of exp.
died            byte    %8.0g                 1 if patient died
drug            byte    %8.0g                 Drug type (1=placebo)
age             byte    %8.0g                 Patient´s age at start of exp.
-------------------------------------------------------------------------------
Sorted by:
```

このサバイバルデータでは生存時間が studytime, イベント発生は died が 1, センサードの場合は 0 です. drug の 1 はプラセボでそれを含めて 3 種類の薬を使っています. ここで推定するモデルは共変量 age の係数は drug に関係なく同一である, という制約をかけたパラメトリックハザードモデルです.

マルチプルグループモデルのフィット

これからフィットするモデルを次に示します.



drug に関する制約を課さないのであれば, 次のコマンドで推定できます.

```
.  gsem (studytime <- age, family(weibull, failure(died)))
```

(推定結果は省略します)

次に drug でグループ分けして, age の係数だけは同一であるとする制約をかけることにします. 従って, 先のコマンドに group(drug) と ginvariant(coef) のオプションを追加し, ハザード関数としてはワイブル分布を利用します.

```
. gsem (studytime <- age, family(weibull, failure(died))),
> group(drug) ginvariant(coef)
```

```
Generalized structural equation model           Number of obs     =         48
Grouping variable  = drug                       Number of groups  =          3
Log likelihood     = -109.28976

 ( 1)  [studytime]1bn.drug#c.age - [studytime]3.drug#c.age = 0
 ( 2)  [studytime]2.drug#c.age - [studytime]3.drug#c.age = 0

Group          : 1                              Number of obs     =         20
Response       : studytime                      No. of failures   =         19
Family         : Weibull                        Time at risk      =        180
Form           : proportional hazards
Link           : log
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
studytime						
age	.1212332	.0367538	3.30	0.001	.049197	.1932694
_cons	-10.36921	2.341022	-4.43	0.000	-14.95753	-5.780896
/studytime						
ln_p	.4541282	.1715663			.1178645	.7903919

```
Group          : 2                              Number of obs     =         14
Response       : studytime                      No. of failures   =          6
Family         : Weibull                        Time at risk      =        209
Form           : proportional hazards
Link           : log
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
studytime						
age	.1212332	.0367538	3.30	0.001	.049197	.1932694
_cons	-14.93039	3.445179	-4.33	0.000	-21.68282	-8.177965
/studytime						
ln_p	.9413477	.2943728			.3643876	1.518308

```
Group          : 3                              Number of obs     =         14
Response       : studytime                      No. of failures   =          6
Family         : Weibull                        Time at risk      =        355
Form           : proportional hazards
Link           : log
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
studytime						
age	.1212332	.0367538	3.30	0.001	.049197	.1932694
_cons	-14.08495	3.242463	-4.34	0.000	-20.44006	-7.72984
/studytime						
ln_p	.6735495	.369625			-.0509022	1.398001

この結果, age の係数はどのグループのそれも等しく 0.12 となっています. それ以外の係数はすべて異なる値となっています.

Notes:

ここで利用したオプション coef の用法を説明しておきます. モデルのパラメータは次に示すクラスに分けて指定します.

クラスの内容	クラス名
1. 定数項とカットポイント	cons
2. 計測可能な変数の係数	coef
3. 潜在変数の係数	loading
4. 誤差共分散	errvar
5. スケールパラメータ	scale
6. 外生変数の平均	means
7. 外生潜在変数の共分散	covex
8.1-7 のすべて	all
9. 何も指定しない	none

ここでは計測可能な変数 age の係数に, すべて等しいという制約をかけたいのでオプション coef を利用して ginvariant(coef) としました.

表示方法を変更する

ただし, 推定結果が drug のグループごとに繰り返し表示されていますので, 係数を比較するにはやや不便です. このようなケースでは次に示す byparm オプションを利用すると便利です.

```
. gsem,byparm
```

```
Generalized structural equation model           Number of obs     =         48
Grouping variable  = drug                       Number of groups  =          3
Group          : 1                              Number of obs     =         20
Response       : studytime                      No. of failures   =         19
Family         : Weibull                        Time at risk      =        180
Form           : proportional hazards
Link           : log
Group          : 2                              Number of obs     =         14
Response       : studytime                      No. of failures   =          6
Family         : Weibull                        Time at risk      =        209
Form           : proportional hazards
Link           : log
Group          : 3                              Number of obs     =         14
Response       : studytime                      No. of failures   =          6
Family         : Weibull                        Time at risk      =        355
Form           : proportional hazards
Link           : log
Log likelihood     = -109.28976
 ( 1)  [studytime]1bn.drug#c.age - [studytime]3.drug#c.age = 0
 ( 2)  [studytime]2.drug#c.age - [studytime]3.drug#c.age = 0
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
studytime    |
         age |
           1 |   .1212332   .0367538     3.30   0.001      .049197    .1932694
```

```
           2 |   .1212332   .0367538     3.30   0.001      .049197    .1932694
           3 |   .1212332   .0367538     3.30   0.001      .049197    .1932694
             |
       _cons |
           1 |  -10.36921   2.341022    -4.43   0.000    -14.95753   -5.780896
           2 |  -14.93039   3.445179    -4.33   0.000    -21.68282   -8.177965
           3 |  -14.08495   3.242463    -4.34   0.000    -20.44006    -7.72984
-------------+----------------------------------------------------------------
/studytime   |
        ln_p |
           1 |   .4541282   .1715663                      .1178645    .7903919
           2 |   .9413477   .2943728                      .3643876    1.518308
           3 |   .6735495    .369625                      -.0509022    1.398001
------------------------------------------------------------------------------
```

このようにすると係数の比較の簡単です. 定数項の値は一見すると, グループ間で異なっているように見えますが, これをウォルド検定で確かめてみましょう. まずは係数の指定方法を確認するために, 次のコマンドを実行します.

```
. gsem,coeflegend
```

(結果は省略します)

それぞれの定数項は_b[studytime:1.drug] のような形式で指定できることが分かりますので, 次のコマンドで検定を行います.

```
. test _b[studytime:1.drug]=_b[studytime:2.drug]=_b[studytime:3.drug]


 ( 1)  [studytime]1bn.drug - [studytime]2.drug = 0
 ( 2)  [studytime]1bn.drug - [studytime]3.drug = 0
           chi2(  2) =    5.49
         Prob > chi2 =    0.0641
```

有意水準 5% で考えると, すべての定数項が等しいという帰無仮説は棄却できません. そこで, 定数項についても等しいという制約をかけてモデルを推定する場合は次のコマンドを利用します.

```
. gsem (studytime <- age, family(weibull, failure(died))), ///
> group(drug) ginvariant(coef cons)
```

(結果は省略します)

パラメトリックサバイバルモデル

ここで紹介した group オプションですが, 実態としては次に示すパラメトリックサバイバルモデルのコマンド streg で strata オプションを利用する事と全く同じ事です.

```
. stset studytime, failure(died)

. streg age, distribution(weibull) strata(drug)
```

以上